The Visual Object Tracking Challenge Results
VOT2018

The emergence of VOT initiative

„Although tracking itself is by and large a solved problem...“, -- Jianbo Shi & Carlo Tomasi CVPR1994 --

• The **VOT initiative** (February 2013)

• Goal: Establish evaluation standards -> development of trackers

• Four pillars of VOT:
  • Evaluation system
  • Datasets
  • Evaluation methodology
  • Community building (VOT challenges)
The VOT challenge evolution

<table>
<thead>
<tr>
<th>Perf. Measures</th>
<th>Dataset size</th>
<th>Target box</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOT2013</td>
<td>ranks, A, R</td>
<td>16, manual select. manual VOT-ST</td>
</tr>
<tr>
<td>VOT2015</td>
<td>EAO, A, R, EFO</td>
<td>60, fully auto manual VOT-ST, VOT-TIR</td>
</tr>
<tr>
<td>VOT2016</td>
<td>EAO, A, R, EFO</td>
<td>60, fully auto manual VOT-ST, VOT-TIR,</td>
</tr>
<tr>
<td>VOT2017</td>
<td>EAO, A, R, EAO_{realtime}</td>
<td>60x2, fully auto auto VOT-ST, VOT-RT, VOT-TIR</td>
</tr>
</tbody>
</table>

• Carefully developed datasets, annotation, measures, toolkits, subchallenges
• VOT2018:
  • Short-term tracking challenge
  • Short-term tracking real-time challenge
  • Long-term tracking challenge
Outline

1. Scope of the VOT2018 (sub) challenges

2. VOT2018 results overview

3. Winner announcement
VOT2018 short-term (ST) challenge

- Short-term, single-target, causal trackers
- Tracker reports the target state as a rotated bounding box

- No re-detection: drift is considered a failure and tracker is reset
The VOT2018 ST dataset

- VOT2017 dataset did not saturate → same dataset used in VOT2018
- Public dataset (60 sequences) + Sequestered dataset (60 sequences)
- Each image annotated by 6 attributes:
  Occlusion, Illumination change, Object motion, Object size change, Camera motion, Unassigned
- Rotated bounding box automatically computed from pre-segmented image
The VOT2018 ST evaluation methodology

• Two weakly correlated measures\(^2\) chosen according to\(^1\):
  • **Robustness** (number of times a is reinitialized)
  • **Accuracy** (average overlap while tracking)
  • + Combination of basic measures (EAO)

\(^1\)Čehovin, Leonardis, Kristan. *Visual object tracking performance measures revisited*, IEEETIP 2016
\(^2\)Kristan et al., A Novel Performance Evaluation Methodology for Single-Target Trackers, IEEETPAMI 2016
The VOT2018 ST real-time challenge

• Required to process sequences at ~20 fps

• The VOT2018 ST public dataset used for this

• Same performance evaluation protocol and measures as VOT2018 ST
The VOT 2018 workshop

VOT2018 LONG-TERM TRACKING CHALLENGE
Long-term tracking (LTT)

• Required long-term tracker properties:
  • Determine when the target has been lost (or disappeared)
  • Re-detect the target after losing the target
## Short-term vs long-term spectrum

<table>
<thead>
<tr>
<th>ST/LT levels</th>
<th>Position reported</th>
<th>Determines target lost?</th>
<th>Target re-detection</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ST\textsubscript{0}: Basic ST</strong></td>
<td>each frame</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td><strong>ST\textsubscript{1}: Basic ST with conservative updating</strong></td>
<td>each frame</td>
<td>not explicitly, selective update of visual model</td>
<td>no</td>
</tr>
<tr>
<td><strong>LT\textsubscript{0}: Pseudo LT</strong></td>
<td>only when visible</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td><strong>LT\textsubscript{1}: Re-detecting LT</strong></td>
<td>only when visible</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

- ST\textsubscript{0} (e.g., KCF\textsuperscript{2}, MS\textsuperscript{3})
- ST\textsubscript{1} (e.g., MDNet\textsuperscript{4}, ECO\textsuperscript{5}) -> easily converted to LT\textsubscript{0}
- LT\textsubscript{1} (e.g., TLD\textsuperscript{5})

\textsuperscript{1}Lukežič, Čehovin, Vojir, Matas, Kristan, *Now you see me: evaluating performance in long-term visual tracking*, arXiv2018

\textsuperscript{2}Enriques et al. PAMI 2015 ; \textsuperscript{3}Comaniciu et al. PAMI 2002; \textsuperscript{4}Nam et al. CVPR2016;

\textsuperscript{5}Danelljan et al. CVPR2017; \textsuperscript{5}Kalal et al. PAMI 2011
VOT2018 LT tracking dataset

- VOT approach: Keep it **sufficiently small**, well annotated and diverse
- Most challenging element: lots of target disappearances
- The dataset that meets these requirements: LTB35$^1$

\[
\text{LTB35} = \\
20 \text{ (from UAV20L$^2$ -- small objects & many disappearances)} \\
+3 \text{ (from$^3$ -- challenging long sequences)} \\
+6 \text{(new from Youtube -- many disappearances)} \\
+6 \text{ (new generated from$^4$)}
\]

---

$^1$Lukežič, et al., Now you see me: evaluating performance in long-term visual tracking, Arxiv2018  
$^2$Mueller et al., A benchmark and simulator for uav tracking, ECCV2016  
$^3$Kalal et al., Tracking-Learning-Detection, TPAMI2010  
VOT2018 LT tracking dataset

- **35 sequences** (146,847 frames)
- **Axis aligned bounding box** annotations (persons, car, motorcycle, bicycle, boat, animals, etc.)
- **Resolution**: 290x217 - 1280x720
- **Average per sequence disappearance**: 12
- **Average target absence period**: 40 frames
- **Nine per-sequence attributes**:
  1. full occlusion
  2. out-of-view motion
  3. partial occlusion
  4. camera motion
  5. fast motion
  6. scale change
  7. aspect ratio change
  8. viewpoint change
  9. similar objects
LT performance measure design

• Requirements: (i) localization accuracy, (ii) target absence prediction accuracy, (iii) re-detection accuracy

• Precision ($Pr$) ... % of all predictions $A_t$ that agree with GT $G_t$

• Recall ($Re$) ... % of all GT boxes that that agree with predictions $A_t$

• F-measure ... a standard Pr/Re tradeoff

$$F = \frac{2PrRe}{(Pr + Re)}$$

1Lukežič, et al., Now you see me: evaluating performance in long-term visual tracking, Arxiv2018
**LT performance measure design**

- Agreement = sufficient overlap:
  \[ \Omega(A_t, G_t) \geq \tau_\Omega \]
  \[ \Omega(A_t(\tau_\theta), G_t) \geq \tau_\Omega \]

- Precision and Recall depend on two thresholds: \( Pr(\tau_\theta, \tau_\Omega), Re(\tau_\theta, \tau_\Omega) \)

- The overlap threshold is avoided by integrating it out

\[
Pr(\tau_\theta) = \int_0^1 Pr(\tau_\theta, \tau_\Omega) d\tau_\Omega = \frac{1}{N_p} \sum_{t \in \{t: A_t(\tau_\theta) \neq \emptyset\}} \Omega(A_t(\tau_\theta), G_t),
\]

\[
Re(\tau_\theta) = \int_0^1 Re(\tau_\theta, \tau_\Omega) d\tau_\Omega = \frac{1}{N_g} \sum_{t \in \{t: G_t \neq \emptyset\}} \Omega(A_t(\tau_\theta), G_t)
\]
Primary LT performance measures

- **Primary measures** are $\text{Pr}(\tau^*_\theta)$, $\text{Re}(\tau^*_\theta)$ and $F(\tau^*_\theta)$ evaluated at detection certainty threshold that maximizes the tracker F-measure.

  - In short-term setup, $F(\tau^*_\theta)$ reduces to a standard ST measure!

- Primary scores thus **fully avoid** manually setting the thresholds.

---

Matej Kristan (matej.kristan@fri.uni-lj.si)

16/33
VOT2018 participation

- Participants **download** the toolkit and the VOT2018 datasets
- Toolkit automatically performs all experiments
- Submission of raw results + tracker code required
- Top ten trackers re-run by VOT committee on the sequestered dataset
The VOT 2018 workshop

VOT2018 ST CHALLENGE RESULTS
VOT2018: 72 trackers tested

Tracking approach:
- DCF (53%)
- SSVM (4%)
- Siam Net (18%)
- RNN (1%)
- CNN match (6%)
- Flow (11%)
- Mean shift (8%)

ST/LT category:
- ST0 (71%)
- ST1 (25%)
- LT1 (4%)
- UniD (76%)

Motion model:
- highD (2%)
- NCVD (7%)
- RWD (15%)

Generative/discriminative:
- Discriminative (76%)
- Generative (24%)

Holistic/parts:
- Holistic (75%)
- Part-based (25%)
VOT2018 ST results on public dataset

- **Top trackers:** (1) LADCF, (2) MFT, (3) SiamRPN, (4) UPDT, (5) RCO, (6) DRT, (7) DeepSTRCF1, (8) SA_Siam_R, (9) CPT, (10) DLSTpp

- **Tracking approach in top 10:**
  8 DCF, except SiamPRN and SA_Siam_R (based on Siamese nets)

- **Features:**
  - All CNN (+handcrafted)
    Top two apply Resnet50
  - Most CNN trained for detection
  - Some trained for localization,
    (e.g. MFT)
VOT2018 results on public dataset

- Top trackers are among the most robust trackers:
  1. MFT, 2. LADCF, 3. RCO, 4. UPDT

- Top in accuracy:
  1. SiamRPN, 2. SA_Siam_R, 3. FSAN, 4. DLSTpp

- Per-attribute analysis:

<table>
<thead>
<tr>
<th>CM</th>
<th>IC</th>
<th>MC</th>
<th>OC</th>
<th>SC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.49</td>
<td>0.47</td>
<td>0.47</td>
<td>0.40</td>
<td>0.43</td>
</tr>
<tr>
<td>0.74</td>
<td>1.05</td>
<td>0.87</td>
<td>1.19</td>
<td>0.61</td>
</tr>
</tbody>
</table>

- Most failures due to: Occlusion
- Mostly affects accuracy: Occlusion + Scale change
VOT2018 results on public dataset

- Baselines ranked at the very tail of the benchmark
- 19 trackers published at major CV venues (2017≤)
  - Their average performance: VOT2018 sota bound
  - Over 26% submissions exceed this bound
VOT2018 results on sequestered dataset

- Fairly stable ranks
- Greatest change: SiamRPN
- Smallest rank change: MFT
VOT2018 ST realtime results

- Top 10: (1) SiamRPN, (2) SA_Siam_R, (3) SA_Siam_P, (4) SiamVGG, (5) CSRTPP, (6) LWDNTm, (7) LWDNTthi, (8) CSTEM, (9) MBSiam, (10) UpdateNet

Two classes:
- Approach: Siamese (e.g., SiamFc\(^1\))
  - Features: CNN
  - Hardware: GPU
- Approach: CSRDCF\(^2\)
  - Features: HOG+CN
  - Hardware: CPU

1Bertinetto, et al., VOT2016, 2Lukežič, et. al., CVPR2017
VOT2018 ST Realtime vs Baseline results

- Most of the top baseline performers drop with real-time constraint
- The drop is smaller for real-time trackers on baseline challenge
- Some achieve top real-time performance AND perform well on the baseline
The VOT 2018 workshop

VOT LONG-TERM CHALLENGE RESULTS
VOT2018 long-term challenge: trackers tested

- 15 trackers tested
- 5 trackers from $ST_0$ class (elevated to $LT_0$)
- 10 trackers from $LT_1$ class
- LT trackers composed of: a short-term component and a detector
Image-wide re-detection test

- Designed to test a crucial ingredient: *Image-wide re-detection*
- Created artificial sequences: *Only target position changes*

Image-wide detection: 8 trackers
(nearly instant detection, with exceptions)

<table>
<thead>
<tr>
<th>Tracker</th>
<th>Frames (Success)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. MBMD</td>
<td>1 (100%)</td>
</tr>
<tr>
<td>2. DaSiam_LT</td>
<td>- (0%)</td>
</tr>
<tr>
<td>3. MMLT</td>
<td>0 (100%)</td>
</tr>
<tr>
<td>4. LTSINT</td>
<td>2 (100%)</td>
</tr>
<tr>
<td>5. SYT</td>
<td>0 (43%)</td>
</tr>
<tr>
<td>6. PTAVplus</td>
<td>0 (11%)</td>
</tr>
<tr>
<td>7. FuCoLoT</td>
<td>78 (97%)</td>
</tr>
<tr>
<td>8. SiamVGG</td>
<td>- (0%)</td>
</tr>
<tr>
<td>9. SLT</td>
<td>0 (100%)</td>
</tr>
<tr>
<td>10. SiamFC</td>
<td>- (0%)</td>
</tr>
<tr>
<td>11. SiamFCDet</td>
<td>0 (83%)</td>
</tr>
<tr>
<td>12. HMMTxD</td>
<td>3 (91%)</td>
</tr>
<tr>
<td>13. SAPKLTF</td>
<td>- (0%)</td>
</tr>
<tr>
<td>14. ASMS</td>
<td>- (0%)</td>
</tr>
<tr>
<td>15. FoT</td>
<td>0 (6%)</td>
</tr>
</tbody>
</table>

7 trackers id not pass the LT1 test: Local detection at most
Overall results – the benchmark

- Top 10: 9 CNN-based (mostly Siamese architectures), 2 apply DCF
- Best: MBMD (Based on MDNet\(^1\)) – bounding box regression + classification
- All top trackers are LT\(_1\) with image-wide detection except DaSiam_LT

\(^1\)Nam, Han, Learning, Multi-Domain Convolutional Neural Networks for Visual Tracking, CVPR2016
Architecture analysis:

- Long-term component (Detector):
  - **CNNs** consistently deliver top performance
  - **DCFs** show great promise (when properly learned)
  - **Deep features** dominate

- Short-term component:
  - Follow trend observed in ST tracking benchmarks
Attribute analysis

Most challenging:
- Fast motion
- Out of view
- Aspect ratio change
VOT2018 challenges Summary

• VOT2018 ST baseline:
  • DCF the dominant methodology, Deep features the dominant features
  • Wider use of deep features trained for localization

• VOT2018 ST realtime:
  • Best performance: fully convolutional deep approaches and DCFs
  • Some of the fastest trackers also rank among best on baseline

• VOT2018 LT:
  • Explicit object detection integrated
  • Nearly all top-performers CNN-based, only one purely DCF
  • Top two performers do not update the detector
VOT2018 online resources

• Results in 155 coauthor paper (58 institutions)
• Available at http://www.votchallenge.net/vot2018
• Presentations + papers + Dataset + Evaluation kit
• Guidelines on how to evaluate your trackers on VOT2018 and produce graphs for your papers (directly comparable to 72 ST and ~10 LT trackers!)

• VOT is open source:
  • All toolkits and protocols on Github
  • Trackers code available:
    VOT2018 (72 ST, 15 LT), VOT2017 (36), VOT2016 (39)
Winners of the VOT2018 short-term challenge:

MFT by: S. Bai, Z. He, J. Zhuang

“Multi-solution Fusion for Visual Tracking”
VOT2018 awards:

Winners of the VOT2018 ST realtime challenge:

SiamRPN by: Q. Wang, Z. Zhu, B. Li, W. Wu, W. Hu, W. Zou

“High Performance Visual Tracking with Siamese Region Proposal Network”

(talk after the Poster session!)
VOT2018 awards:

Winners of the VOT2018 long-term challenge:

MBMD by: Y. Zhang, L. Wang, D. Wang, J. Qi, H. Lu

“MobileNet based tracking by detection algorithm”

(talk after the Poster session!)
Thanks

• The VOT2018 committee

• Everyone who participated or contributed


• VOT2018 sponsors:

Matej Kristan (matej.kristan@fri.uni-lj.si)